Multi-Modal 기반 음성, 텍스트 연관성 학습을 적용한 자유발화 감정 예측 알고리즘

Free Speech Emotion Prediction Algorithm by Applying Multi-Modal Based Voice and Text Correlation Learning

요 약

최근 COVID-19로 인해 인간은 비대면과 개인화된 생활에 익숙해졌다. 이러한 상황은 사람들에게 소통할 수 있는 기회를 제한하였고, 사회성 발달과 인격 형성, 공감능력에 부정적 영향을 보였다. 팬데믹 전으로 돌아가고 있는 요즘, 이러한 상황은 원활한 의사소통에 부정적 영향을 끼칠 수 있으며, 대화에서의 감정 분석은 원활한 의사소통에 중요한 영향을 줄 수 있다. 이러한 문제를 해결하기 위해 본 논문에서는 Multi-Modal 기반 음성, 텍스트 연관성을 학습하여 다중 감정 예측 방법을 제안한다. 먼저, 자유발화의 음성 정보와 텍스트 정보에 대해서 전처리를 진행한다. BERT 알고리즘의 WAV-Transformer, TXT-Transformer를 결합하여 7가지의 다중 감정(기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔)을 예측한다. 제안한 모델의 성능을 평가하기 위해 베이스라인인 CNN, LSTM, CNN-LSTM과 성능을 비교하였으며, 제안한 모델이 더 잘 예측함을 보였다.

1. 서 론

최근 IT 기술이 크게 발전함에 따라, 기술이 음성 인식, 감정 분석 등 다양한 분야에 적용[1]되고 있다. 또한 COVID-19로 인한 대화의 단절, 업무의 패턴 변화로 일상 대화의 감정을 이해하는 것은 매우 어려운 일이다. 만약 대화에서의 감정을 미리 파악할 수 있다면 원활한 의사소통에 큰 도움이 될 것이다.

이와 관련된 연구에서, Bang et al.[2]은 MLP-Mixer 구조를 활용하여 Multi-Modal 감정 예측 방법을 제안했다. Ko et al.[3]은 GAN(Generative Adversarial Network) 알고리즘을 이용하여 부정 감정 데이터를 생성함으로써 음성 기반 감정 인식 모델의 성능을 개선하였다. 그러나, 이러한 연구들은 다양한 정보와 데이터를 사용하여 모델을 검증해야한다.

이러한 문제를 해결하기 위해 본 논문에서는 발화 음성 데이터와 텍스트 데이터를 모두 사용하여 음성-텍스트 기반 자유발화 감정 예측 알고리즘을 제안한다. 자세히는, 자유발화 음성-텍스트 데이터의 연관성에 대한 정보를 추출한다. 추출한 정보를 BERT 알고리즘[4]으로 모델을 학습한다. 최종적으로 대화에 대한 감정(기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔) 을 예측한다. 제안한 모델의 성능을 평가하기 위해 본 논문에서는 동일한 데이터를 활용하여 베이스라인인 CNN[4], LSTM[4], CNN-LSTM[4]과 성능을 비교하였고 감정을 더 잘 예측하는 것을 보였다. 추가적으로 통계검증[4]을 통해 제안한 모델이 더 유의미한 차이를 보였다.

본 논문의 기여도는 다음과 같다.

- 음성-텍스트 기반 자유발화를 사용하여 BERT 모델을 학습하였으며, 독립된 음성, 텍스트 기반 모델보다 음성과 텍스트를 결합한 모델이 성능을 더 개선하였다.
- 발화 음성-텍스트 데이터에 대한 연관성 추출을 사용하여 BERT 모델을 학습하였으며, 베이스라인보다 더 잘 예측함을 보였다.

만약 자유발화에서의 감정 분석을 자동적으로 진행할 수 있다면, 원활한 의사소통에 도움을 줄 수 있다.

2. 관련연구

딥러닝 알고리즘의 다양한 적용으로, 다중 감정 인식에 대한 연구가 활발하게 진행되고 있다. Zhao et al.[5]는 Wav2vec와 BERT 알고리즘을 활용하여 다중 감정 인식을 예측하였다. Jeong et al.[6]은 그래프 구조 기반 대화 정보를 추출하여 대화의 감정을 예측하였다. Graph Attention을 사용하여 모델을 제안하여 베이스라인과 비교 실험을 진행하였다. 제안한 모델의 성능이 감정 분류에서 기존보다 높은 성능을 보였다. Xia et al.[7]은 잠재적인 감정 쌍(Pair)과 해당 원인을 추출하는 것을 목표로 하는 감정-원인 쌍 추출(ECPE) 방법을 제안하였다. Pepino et al.[8]은 사전 훈련된 Wav2vec 2.0 모델에서 추출한 특징을 얕은 신경망의 입력으로 사용하여 음성의 감정을 인식하는 음성 감정 인식 학습 방법을 제안했다. 사전 훈련된 모델에서 여러 레이어의 출력을 결합하여 풍부한 음성 표현을 생성하는 방법을 제안하였다.

3. 본 론

Multi-Modal 기반 다중 감정 예측을 위한 모델의

전반적인 도식도는 다음 그림 1과 같다.

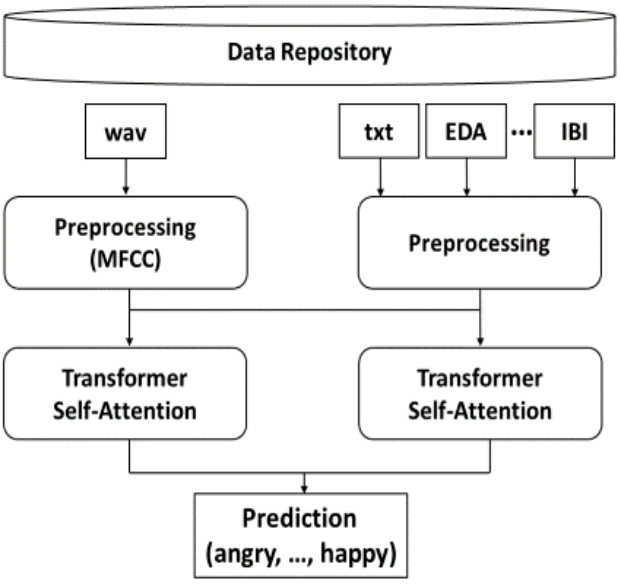

그림 1. 제안한 전반적인 도식도

자유발화에서의 정보(WAV, TXT, Temperature, EDA, IBI)를 활용하여 음성에 대해서는 MFCC 알고리즘을 사용하고, 그 외 정보는 전처리 과정을 진행한다. 추출한 특징을 BERT 알고리즘의 입력으로 학습하고, 최종적으로 다중 감정을 예측한다.

3.1 전처리

본 논문에서는 음성, 텍스트 데이터를 BERT 모델에 학습하기 위해 데이터 전처리 과정을 진행한다. 음성 데이터에 대해서는 MFCC 알고리즘[5]을 사용한다. MFCC는 음성인식 및 음성처리 분야에서 널리 사용되는 특징 추출 알고리즘이다. 음성신호를 Mel 스케일에서 Mel 주파수 분석을 수행하고 이후에 이산 코사인 변환을 적용하여 주파수 스펙트럼을 적은 수의 코사인 계수로 표현한다. 이 과정을 통해 음성 신호의 주파수 특징과 유용한 정보를 추출한다. Text 데이터는 특수기호를 제거하고 형태소 분석을 통해 '명사'의 단어를 추출하여 정보를 추출한다.

3.2 BERT 알고리즘

제안한 BERT 알고리즘의 구조는 다음과 같다. 음성과 텍스트 정보에 대해 각각의 Transformer 인코더가 존재하고, 각 인코더는 Self-Attention과 Feed-forward Neural Network로 구성한다. 마지막으로, 각 인코더의 출력은 두개의 Fully Connected Layer를 진행하여, 다중 감정을 예측한다.

4. 실 험

4.1. 데이터

제안한 모델의 실험을 진행하기 위해 공개된 Multi-Modal 감정 데이터[9]를 사용한다. 이 데이터는 한국어 기반 Multi-Modal 감정 데이터셋으로, 음성과 음성의 텍스트 내용, EDA, 신체 온도 데이터로 구성된다. 감정 분류는 크게 7가지(기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔)로 분류된다. 추가적으로, 본 논문에서는 음성 파일은 있지만 음성 데이터가 포함되어 있지 않은 데이터는 결측치로 판단하여 실험에서 제외한다.

4.2 평가 척도

모델의 효율성을 평가하기 위해 본 논문에서는 머신러닝/딥러닝 분야에서 자주 사용하는 Precision, Recall, F-Measure 평가 척도[4]를 사용한다. 데이터는 학습 (8) / 검증 (1) / 평가 (1) 데이터로 구성하며 평가를 진행하고, 데이터의 편향을 줄이기 위해 K-Fold 교차검증[4](K=10)을 진행한다.

4.3 베이스라인

본 논문의 모델을 비교 평가하기 위해 다음과 같이 베이스라인을 설정한다.

(1) CNN[4]: 인공 신경망 모델로, 입력 정보에 대하여 합성곱 연산과 풀링 연산을 진행한다. 그리고 추출한 특징을 사용하여 모델을 학습하고 예측한다.
(2) LSTM[4]: 입력 데이터와 이전 출력 데이터를 통해 현재의 출력을 계산하는 모델이다. Input/Output/Forget 게이트를 사용하여 입력된 데이터를 처리하고, 어떠한 정보를 특정 비율로 유지할 것인지 결정한다.
(3) CNN-LSTM[4]: CNN과 LSTM을 결합하여 시퀀스 데이터를 처리하는 모델이다. CNN을 이용하여 주요한 특징을 추출하고, LSTM에서 이러한 정보를 사용하여 시퀀스를 분석한다.

4.4 실험 결과

제안한 모델의 Multi-Modal 감정 예측 성능 결과는 다음 그림 2와 같다. BERT(TXT)는 대화 텍스트 정보만 활용했을 때를 의미하고, BERT(Wav)는 대화 음성 정보만 활용했을 때를 의미한다. BERT(TXT+WAV)는 대화 텍스트와 음성 정보를 활용했을 때를 나타낸다.

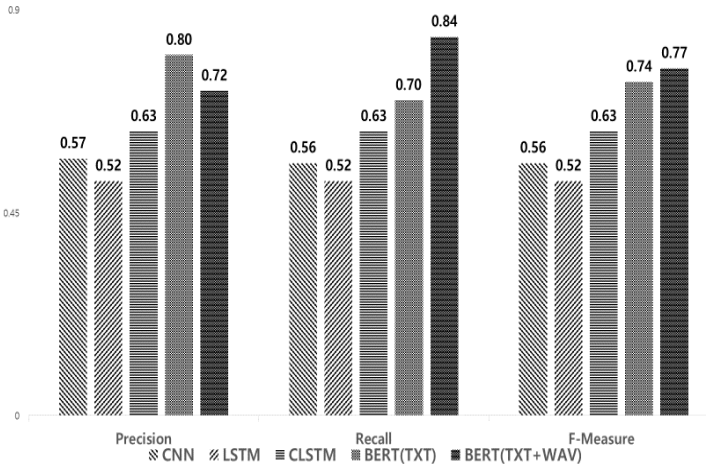

그림 2. 제안한 다중 감정 예측 성능 결과

X축은 평가척도(Precision, Recall, F-Measure)를 의미하고, Y축은 값(Value)을 나타낸다. 제안한 모델은 Precision에서 0.72, Recall에서 0.84, F-Measure에서 약 0.77의 성능을 보였다. 전반적으로 베이스라인 보다 다중 감정을 잘 예측함을 보인다.

추가적으로, 제안한 모델이 베이스라인보다 유의미한 차이가 있음을 보이기 위해 통계 검증[4]을 진행한다. 설정한 귀무가설(H_0)과 대립가설(H_a)은 다음과 같다.

- $H1_0$, $H2_0$, $H3_0$: 제안한 모델은 CNN, LSTM, CLSTM 보다 유의미한 차이가 없다.
- $H1_a$, $H2_a$, $H3_a$: 제안한 모델은 CNN, LSTM, CLSTM 보다 유의미한 차이가 있다.

통계 검증을 진행하기 위해 먼저, 제안한 모델과, 베이스라인의 F-Measure 값에 대해 정규성 검증[4]을 진행한다. 정규성 검증 결과 0.05보다 크면 T-검증[4]을 진행하고, 작으면 Wilcox-검증[4]을 진행한다.

통계 검증 결과는 다음 표 1과 같다.

표1. 통계 검증 결과

Null Hypothesis	p-value	Result
$H1_0$	2.45E-10	$H1_a$: Accept
$H2_0$	2.06E-11	$H2_a$: Accept
$H3_0$	2.21E-06	$H3_a$: Accept

H1은 제안한 모델과 CNN 알고리즘간 유의미한 차이가 없음을 나타낸다. 통계 검증 결과 H1의 p-value는 2.45E-10(0.000000000245)이고 0.05보다 작으므로 귀무가설을 기각하고 대립가설을 승인한다. 따라서 제안한 모델은 CNN 알고리즘간 유의미한 차이가 있음을 보이고, 모든 귀무가설을 기각하고, 대립가설은 승인한다.

5. 토 의

본 논문의 실험 결과에서 BERT(TXT+WAV) 모델은 약 77%의 F-Measure을 보였으며, 전반적으로 다중 감정을 잘 예측함을 보였다. 추가적으로, 제안한 모델이 다른 모델보다 통계 검증에서 유의미한 차이가 있음을 보였다.

- 자유 발화 음성과 텍스트 연관성을 학습하여 모델을 훈련하였으며, 베이스라인보다 성능이 더 개선되었다.
- 텍스트 데이터만 사용했을 때, CNN, LSTM, CLSTM과 비교하였으며, 제안한 BERT(TXT) 모델이 더 잘 예측함을 보였다. 그리고 BERT(TXT+WAV)는 BERT(TXT) 보다 더 성능을 개선하였다.

6. 결 론

일상 대화에서의 감정 예측은 원활한 대화를 진행하는데 중요한 역할을 할 수 있다. 본 논문에서는 Multi-Modal 기반 자유발화 음성과 텍스트 정보를 활용하여 다중 감정 예측 알고리즘을 제안했다. 먼저, 음성, 텍스트 연관성에 대해 특징을 추출하고, 데이터 전처리를 진행한다. 다음, BERT 알고리즘을 활용하여 모델을 학습하고, 최종적으로 해당 자유발화에 대한 7가지 다중 감정을 예측했다. 제안한 s모델의 성능을 평가하기 위해 본 논문에서는 베이스라인(CNN, LSTM, CNN-LSTM)과 비교를 진행하였고, 제안한 모델이 다중 감정을 더 잘 예측한 것을 보였다. 추가적으로 통계 검증을 통하여 제안한 모델이 베이스라인보다 더 유의미한 차이가 있음을 검증하였다. 향후 다양한 데이터와 추가 특징 정보를 활용하여 모델을 더 확장할 계획이다.

7. 참고문헌

[1] S., Lee, Y., Yoon, G., Lee, and J. Joh, "Development of Speech Recognition Emotion Analysis Program using Machine Learning", In Journal of 한국컴퓨터 교육학회, 학술발표대회 논문집, 22(2), pp 71-73, 2018.

[2] N., Bang, H., Yeen, J., Lee, and M., Koo, “MMM: Multi-modal Emotion Recognition in Conversation with MLP-Mixer”, In Proc. of 2022년 한국 소프트웨어 종합 학술 대회 논문, pp.2288-2290, 2022.

[3] Y., Ko, and Y., Kim, "Performance Improvement of Speech Emotion Recognition Model Using Generative Adversarial Networks", In Journal of Korean Institute of Information Technology, vol.17 no.11, pp.77-85, 2019.

[4] C., Zhang, M., Cai, X., Zhao, and D., Wang, "Research on Case Preprocessing based on Deep Learning", In Journal of Concurrency Computat Pract Exper, vol. 34, no. 2, 2022.

[5] Z., Zhao, Y., Wang, Y., Wang, “Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition”, In Cooperative Medianet Innovation Center, Shanghai Jiao Tong University Shanghai AI Laboratory, pp.1-5, 2022.

[6] D., Jeong, and J., Bak, “Extracting Emotion-Cause Information From Conversation Data Using The Graph Structure”, In Proc. of 2021년 한국 소프트웨어 종합 학술 대회 논문, pp.515-517, 2021.

[7] R., Xia, and Z., Ding, “Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts”, In Association for Computational Linguistics, pp. 1-10, 2019.

[8] L., Pepino, P., Riera, and L., Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings.", arXiv preprint arXiv:2104.03502, 2021.

[9] ETRI KEMDy20 데이터 2023, “https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR”, accessed April 30, 2023.